Comp. & Maths. with Appls. Vol. 1, pp. 69-78. Pergamon Press, 1975. Printed in Great Britain.

ON DESIGNING SIMULATION MODELS FOR EVALUATING DISCRIMINANT ANALYSIS ROUTINES*

RAJ S. CHHIKARA and PATRICK L. ODELL

The University of Texas at Dallas, Dallas, Texas, U.S.A.

(*Received* 4 *September* 1973)

Abstract—This paper gives certain simulation models that can be used to verify a Bayes' discriminant algorithm and a feature selection routine for their accuracy. The verification is developed in terms of statistical tests and the evaluation techniques are suggested, assuming the multivariate normal model for underlying classes. The algorithm evaluation procedures are discussed from a practical viewpoint and are easy to perform.

1. INTRODUCTION

Consider the statistical problem of classification involving m different classes $\pi_1, \pi_2, \ldots, \pi_m$. Let each individual in these classes possess p common observable features $C_1, C_2, \ldots, C_p$. The observation on an individual is denoted by a $p \times 1$ measurement vector $x = (x_1, x_2, \ldots, x_p)^T$, where x_j denotes the measurement corresponding to feature C_j. Let $p_i(x)$ denote the multivariate probability density function of the random vector x for π_i and q_i be *a priori* probability that an individual be selected from a class π_i, $i = 1, 2, \ldots, m$. Then the statistical classification problem consists of devising a technique for assigning an individual selected at random from the population made of $\pi_1, \pi_2 \ldots, \pi_m$ into one of the m classes. Among several suggested techniques for solving the problem, the Bayes discriminant procedure leads to optimal solution, in the sense that it minimizes the expected cost of misclassification. However, it is generally not possible to find an exact Bayes solution explicitly due to difficulties involved in the evaluation of classification errors[1]. Also, sometime in its application the amount of computation involved could be immense. In view of this certain modifications in the Bayes' discriminant procedure have been suggested and this has led to various Bayes' algorithms, for example, table look-up techniques[2, 3]; hoping that these algorithms would maintain approximately the optimality of the Bayes' solution.

From a computational viewpoint it is sometimes desirable to reduce the magnitude of a discriminant problem. Since the number of classes, m, is not arbitrary the only way to achieve reduction in the problem is to reduce p, the number of features. This amounts to selecting a "best" set of q features, $q < p$, out of the set of p features; by "best" we mean that when a discriminant analysis is performed using the set of q selected features in place of the complete set of p features, the results are well approximated. A list of different feature selection procedures can be found in Levine[4] who has summarized the earlier work on the topic fairly well.

*Research supported by NASA Contract NAS 9-11925.

In the absence of any knowledge on the probability density functions $p_1(x)$, $p_2(x)$, ..., $p_m(x)$, the Bayes algorithms, characterized as sample-based[5], would often result into further loss in optimality of the Bayes' procedure. Thus when a classification task is performed using a sample-based Bayes' discriminant algorithm coupled with a feature selection routine, the process may involve a complicated algorithm.

The above discussion now raises the basic question as how to verify a classification algorithm. In other words, how to find whether or not an algorithm is accomplishing what it was designed for with theoretical precision? An answer to this question can be found by evaluating the classification algorithm by using real data as input. However, often we do not know in exact forms the statistical properties of the real data and so the algorithm must be verified with data from known probability distribution functions.

Most investigators have tested a software package by verifying it through an example for which analytically they already knew the answer. Since the answers to classification problems are in terms of expected classification errors, one should be determining whether or not the algorithm output statistically confirms such expected results. As the Monte Carlo technique can often be employed for estimation purposes, one must devise simulation models for which the algorithm output can be compared with the known *expected* results.

In this paper we consider certain specified situations and formulate simulation models that can be used to verify a discriminant algorithm and a feature selection routine. Evaluation procedures are given by assuming class models to be multivariate normal with some simplified structure for their mean vectors and covariance matrices. Subsequent to our discussion actual simulation results evaluating different classification algorithms, though omitted here, can be given easily.

2. NOTATIONS AND PRELIMINARIES

Let $C(i|j)$ be the cost of misclassifying an individual from class π_j into class π_i and $R = (R_i, R_2, \ldots, R_m)$ be a classification rule such that if an observation x belongs to region R_i, we assign it to the class π_i, $i = 1, 2, \ldots, m$. Then the expected cost of misclassification for an individual under classification rule R is

$$L(R) = \sum_{j=1}^{m} q_j \sum_{\substack{i=1 \\ i \neq j}}^{m} C(i|j)P(i|j), \tag{1}$$

where

$$P(i|j) = \int_{R_i} p_j(x)\,\mathrm{d}x \tag{2}$$

is the probability of misclassifying the individual from π_j into π_i. It is well-known[6] that for a given cost, $C(i|j)$'s, and given *a priori* probabilities, q_i's, the Bayes procedure minimizes equation (1) and it assigns x to R_j if

$$\sum_{\substack{i=1 \\ i \neq j}}^{m} q_i p_i(x) C(j|i) \leq \sum_{\substack{i=1 \\ i \neq k}}^{m} q_i p_i(x) C(k|i), \qquad k = 1, 2, \ldots, m. \tag{3}$$

However, we would assume equal costs of misclassification and zero cost of correct classification, i.e. $C(i|j) = C$ for all i and j, $i \neq j$ and $C(i|i) = 0$ for all i. Then the Bayes' procedure minimizes the expected probability of misclassification and the associated discriminant region R_j consists of x's satisfying

$$q_j p_j(x) = \max_{i=1,\ldots,m} [q_i p_i(x)], \tag{4}$$

$j = 1, 2, \ldots, m$. For when q_i's are all equal, equation (4) reduces to

$$p_j(x) = \max_{i=1,\ldots,m} [p_i(x)] \tag{5}$$

and this is equivalent to having a maximum likelihood solution for the discriminant problem.

When the probability distribution for a class is assumed to be multivariate normal, the density function of a $p \times 1$ random vector is of the form:

$$p(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\tfrac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right] \tag{6}$$

where μ is the mean vector and Σ is the covariance matrix. To generate a random vector x with this distribution, first generate a $p \times 1$ vector y of independent normal components with zero means and then obtain x using the relation

$$x = B^{-1}y + \mu$$

where B is given by $\Sigma^{-1} = B^T B$. When Σ is a non-singular matrix, B can be easily obtained as a unique upper triangular matrix computed using a Cholesky matrix square root algorithm. If a large number of these normal vectors are required to be simulated it is recommended to use the procedure[7] that minimizes the computer time. The techniques for generating normal random numbers are well known[8].

3. SIMULATION MODELS FOR EVALUATING DISCRIMINANT ALGORITHMS

To evaluate a discriminant algorithm for its accuracy a natural way is to ascertain whether or not classification errors (probabilities of misclassification) obtained from the algorithm in a specified situation agree with the expected ones. If not, and the two differ beyond a certain *level of significance*, the algorithm should be recognized inadequate. The test procedure of goodness of fit using the chi-square test, or the like, may be used for this purpose.

When dealing with discriminant analysis for multivariate normal classes, it is often difficult to evaluate expected Bayes' classification errors unless the underlying class distributions are considered to be of a simple form. In view of this, one should design simulation models for which either expected Bayes' classification errors are easily available so that a test of goodness of fit can be used or a test procedure not involving these errors can be given. Below we consider two such cases and formulate simulation models for an algorithm evaluation.

3.1. *Model with colinear and equally spaced mean vectors*

Considering multivariate normal classes with colinear and equally spaced mean vectors and circular ellipsoids of concentrations, let $\mu_i = (ir, 0, 0, \ldots, 0)^T$ and $\Sigma_i = kI$, where r and k are some positive constants, for π_i, $i = 1, 2, \ldots, m$. Clearly the p components of the random vector $x = (x_1, x_2, \ldots, x_p)^T$ are independently distributed and the p-dimensional problem reduces to one-dimensional as all the discriminating information is in the first component x_1 of the random vector x.

For a given *a priori* probabilities $q_1, q_2, \ldots, q_m$, it follows from equation (4) that the Bayes' discriminant region R_i is given by x's satisfying

$$\left[x_1 - \frac{r}{2}(i+j)\right]\frac{r}{k}(i-j) \geq \log\frac{q_j}{q_i} \tag{7}$$

$$j = 1, 2, \ldots, m, j \neq i.$$

By solving the simultaneous inequalities in equation (7) for $i = 1, 2, \ldots, m$, the Bayes' discriminant regions $R_1, R_2, \ldots, R_m$ are obtained as

$$R_1 = \left\{x : x_1 \leq \frac{3}{2}r - \frac{k}{r}\log\frac{q_2}{q_1}\right\},$$

$$R_i = \left\{x : \frac{2i-1}{2}r - \frac{k}{r}\log\frac{q_i}{q_i - 1} < x_1 < \frac{2i+1}{2}r - \frac{k}{r}\log\frac{q_{i+1}}{q_i}\right\},$$

$$i = 2, 3, \ldots, m-1,$$

and

$$R_m = \left\{x : \frac{2m-1}{2}r - \frac{k}{r}\log\frac{q_m}{q_{m-1}} < x_1 < \infty\right\}.$$

Accordingly, the expected probabilities of misclassification are

$$P(i|j) = \int_{R_i} \phi(jr, k)\, dx_1, i \quad \text{and} \quad j = 1, 2, \ldots, m, i \neq j,$$

where $\phi(\mu, k)$ denotes the univariate normal density function with mean μ and variance k. Equivalently,

$$P(1|j) = \Phi\left([\tfrac{3}{2} - j]\frac{r}{\sqrt{k}} - \frac{\sqrt{k}}{r}\log\frac{q_2}{q_1}\right),$$

$$P(i|j) = \Phi\left([i - j + \tfrac{1}{2}]\frac{r}{\sqrt{k}} - \frac{\sqrt{k}}{r}\log\frac{q_{i+1}}{q_i}\right) - \Phi\left([i - j - \tfrac{1}{2}]\frac{r}{\sqrt{k}} - \frac{\sqrt{k}}{r}\log\frac{q_i}{q_{i-1}}\right),$$

$$i = 2, 3, \ldots, m-1$$

and

$$P(m|j) = 1 - \Phi\left([m - j - \tfrac{1}{2}]\frac{r}{\sqrt{k}} - \frac{\sqrt{k}}{r}\log\frac{q_m}{q_{m-1}}\right),$$

$j = 1, 2, \ldots, m$, where $\Phi(a)$ denotes the cumulative standard normal distribution function at point a.

For fixed r and k, the expected Bayes' classification errors, $P(i|j)$'s, are now available and if desired, one can obtain the probability of misclassifying an individual to π_i by calculating

$$f_i = \sum_{\substack{j=1 \\ j \neq i}}^{m} q_j P(i|j).$$

Next, if *a priori* probabilities are assumed equal, the expressions for $P(i|j)$'s are further simplified greatly.

On the other hand, let $N(i|j)$ denote the number of times the Bayes algorithm, which is to be evaluated, incorrectly assigned observations from π_j into π_i in a simulated data of N_j observations for π_j in the above model. Then

$$\hat{P}(i|j) = N(i|j)/N_j,$$

the observed proportion of observations from π_j classified into π_i provides an unbiased estimate of $P(i|j)$. Now if $P(i|j) > 0$ for all i and j such that for any i and j, $N_jP(i|j)$ is not too small, say less than 5, using the goodness of fit criterion one can test the hypothesis of the algorithm being accurate by calculating

$$\chi^2 = \sum_{j=1}^{m} \sum_{i=1}^{m} \frac{[P(i|j) - \hat{P}(i|j)]^2}{P(i|j)}$$

and comparing it with $\chi^2_\alpha(m^2 - m)$, the value of χ^2 variate with $m^2 - m$ degrees of freedom at a given significance level α. If the calculated χ^2 exceeds $\chi^2_\alpha(m^2 - m)$, one should suspect the underlying algorithm procedure.

3.2. *Symmetrically designed models*

The simulation evaluation model in Section 3.1 is based on a highly restrictive configurated design. We would now consider another model which is less restrictive and provides an alternative way to evaluate a discriminant algorithm when expected classification errors cannot be obtained.

First we formulate the concept as follows:

Definition 1: A simulation model D is a system consisting of a set of density functions $p_1(x)$, $p_2(x), \ldots, p_m(x)$ with mean vectors $\mu_1, \mu_2, \ldots, \mu_m$ and covariance matrices Σ_1, $\Sigma_2, \ldots, \Sigma_m$, respectively.

Definition 2: A model is symmetric if the density function with mean vector μ_i is in D, then the density function with mean vector $-\mu_i$ is also in D and the two such densities are equally likely and have the same covariance matrix.

To construct a symmetric model, we consider the set of admissible mean vectors $\mu = (\mu_{(1)}, \mu_{(2)}, \ldots, \mu_{(p)})^T$ defined by

$$S_p = \{\mu : \mu_{(i)} \in \{1, 0, -1\} \text{ except } \mu \text{ being a zero vector}\}.$$

The set S_p consists of $3^p - 1$ vectors. If we denote a vector having 1 for the ith component and 0 for the other $p - 1$ components by e_i, then S_p is the set of all possible lattice points generated by the p vectors e_i's, and if $\mu \in S_p$, then $-\mu \in S_p$. Accordingly, elements in S_p form the following sequence of vectors:

$$\{e_1, -e_1, e_2, -e_2, \ldots, e_p, -e_p, (e_1 \pm e_2), -(e_1 \pm e_2), \ldots, (e_1 + e_2 + \ldots + e_p),$$
$$-(e_1 + e_2 + \ldots + e_p)\}.$$

Assuming $m \le 3^p - 1$ and m even, select the class mean vectors from the sequence $\{re_1, -re_1, \ldots, r(e_1 \pm e_2), -r(e_1 \pm e_2), \ldots, r(e_1 + e_2 + \ldots + e_p), -r(e_1 + e_2 + \ldots + e_p)\}$ such that

$$\mu_j = -\mu_{m/2+j}, j = 1, 2, \ldots, \frac{m}{2} \tag{8}$$

in order to maintain a symmetry in the pattern of means; and the size for the class means is achieved by introducing a preassigned non-negative constant r. Now considering a set of m density functions with mean vectors obtained according to equation (8), covariance matrices taken according to equation (9),

$$\Sigma_j = \Sigma_{m/2+j}, \qquad j = 1, 2, \ldots, \frac{m}{2}, \tag{9}$$

and assuming $q_j = q_{m/2+j}$, $(j = 1, 2, \ldots, m/2)$, we obtain a symmetric model involving m classes.

It can be easily shown that for expected Bayes' classification errors in the case of a symmetrical model,

$$P(i|j) = P(i'|j') \tag{10}$$

where $i' = (m/2 + i) \bmod m$ and $j' = (m/2 + j) \bmod m$, i and $j = 1, 2, \ldots, m$. In view of this a Bayes' algorithm can be evaluated by finding whether or not its output confirms equation (10) with simulated data for a symmetric model.

Suppose we have m multivariate normal densities for the classes with means $\mu_1, \mu_2, \ldots, \mu_m$ chosen according to equation (8) and covariance matrix $\Sigma_i = kI$ for all i, where k is any positive constant. Then it follows from equation (4) that the Bayes' discriminant regions are given by

$$R_i = \{x : x^T(\mu_i - \mu_j) \geq \tfrac{1}{2}(\mu_i + \mu_j)^T(\mu_i - \mu_j) + k \log\frac{q_j}{q_i},$$
$$j = 1, 2, \ldots, m, \qquad j \neq i\}, \qquad i = 1, 2, \ldots, m.$$

Considering $q_1 = q_2 = \ldots = q_m$ and simplifying further, the inequalities in R_i's are of the form

$$\sum_1^p a_t x_t \geq \frac{r}{2} \sum_1^p a_t b_t$$

where a_t's and b_t's take on values in the set $\{-2, -1, 0, 1, 2\}$. There will be a symmetric pattern among discriminant regions and for the classification errors equation (10) holds. Furthermore, as shown in the following example, the discriminant regions may be simplified greatly and may even lead to an easy evaluation of the expected Bayes' classification errors; in which case a Bayes' algorithm can be evaluated using a test of goodness of fit described in Section 3.1.

Example 1: Let $p = 2$ and $m = 4$. Then the set S_2 consists of the following vectors:

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$
$$\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} -1 \\ -1 \end{bmatrix},$$

Suppose we select the mean vectors

$$\mu_1 = r\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \mu_2 = r\begin{bmatrix} -1 \\ 1 \end{bmatrix}, \qquad \mu_3 = r\begin{bmatrix} -1 \\ -1 \end{bmatrix}, \qquad \mu_4 = r\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

for classes π_1, π_2, π_3 and π_4 respectively. Then the Bayes' discriminant regions R_1, R_2, R_3, R_4 correspond to first, second, third and fourth quadrants of (x_1, x_2)-coordinate plane. (Here we have assumed $q_1 = q_2 = q_3 = q_4$.) One can now easily evaluate expected classification errors $P(i|j)$'s for any specified values of r and k because

$$P(1|1) = P(3|3) = \left[1 - \Phi\left(-\frac{r}{\sqrt{k}}\right)\right]^2$$

$$P(2|2) = P(4|4) = \left[\Phi\left(\frac{r}{\sqrt{k}}\right)\right]\left[1 - \Phi\left(-\frac{r}{\sqrt{k}}\right)\right]$$

$$P(2|3) = P(4|1) = \Phi\left(\frac{r}{\sqrt{k}}\right)\Phi\left(-\frac{r}{\sqrt{k}}\right)$$

$$P(2|4) = P(4|2) = P(1|3) = P(3|1) = \Phi^2\left(-\frac{r}{\sqrt{k}}\right)$$

$$P(1|2) = P(1|4) = P(2|1) = P(3|4) = P(3|2) = P(4|3) = \left[\Phi\left(-\frac{r}{\sqrt{k}}\right)\right]\left[1 - \Phi\left(-\frac{r}{\sqrt{k}}\right)\right].$$

For the case where expected classification errors cannot be determined explicitly, a Bayes' algorithm can be verified by obtaining for it $N(i|j)$, the number of times observations from π_j classified into π_i for simulated data consisting of N_j observations, $j = 1, 2, \ldots, m$ and $i = 1, 2, \ldots, m$, and then testing the null hypothesis that equation (10) holds. Again, the χ^2 test can be used since the random variable

$$\sum_{j=1}^{m} \sum_{i=1}^{m} \frac{[N(i|j) - N_j(N(i|j)) + N(i'|j'))/(N_j + N_{j'})]^2}{N_j[N(i|j) + N(i'|j')]/(N_j + N_{j'})}$$

has an approximate χ^2 distribution with $m(m - 1)/2$ degrees of freedom. If the null hypothesis is rejected at a given α level of significance, it would imply that the algorithm has failed in its mission and needs to be rectified for the inaccuracy and malfunctioning.

4. EVALUATION OF FEATURE SELECTION ROUTINES

A feature selection routine is devised for selecting a characteristic vector $C = (C_1, C_2, \ldots, C_q)^T$ from a larger vector of p characteristics (features) so that the underlying classes are discriminated in the best possible way on the basis of the selected features. A general approach in solving a discriminant problem is to define a distance function $D(i, j)$ between two classes π_i, π_j for each possible characteristic vector C and then select the characteristic vector for which the minimum distance between a pair of classes is maximized. One such distance function is *divergence*[9] defined by

$$D(i, j) = \int_{-\infty}^{\infty} [p_i(x) - p_j(x)] \log [p_i(x)/p_j(x)] \, dx \qquad (11)$$

which has often been considered for the feature selection procedure [4, 10, 11]. Though in general no one-to-one relationship exists between divergence and Bayes' classification errors[12], intuitively one would feel that the expected cost of misclassification decreases with increasing pairwise distances between classes.

Letting $p_i(x)$ be the multivariate normal density of the form (6) with mean vector μ_i and covariance matrix Σ_i, $i = 1, 2, \ldots, m$, we get

$$D(i,j) = \tfrac{1}{2}\{tr(\Sigma_i - \Sigma_j)(\Sigma_j^{-1} - \Sigma_i^{-1}) + (\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_1 - \mu_j)(\mu_i - \mu_j)^T\}. \quad (12)$$

In case of $\Sigma_i = \Sigma$ for all i,

$$D(i,j) = tr\{\Sigma^{-1}(\mu_i - \mu_j)(\mu_i - \mu_j)^T\}. \quad (13)$$

Further, if we let $\Sigma = \text{diag}\ (\sigma_i^2,\ i = 1, 2, \ldots, p)$, a matrix with diagonal elements as σ_1^2, $\sigma_2^2, \ldots, \sigma_p^2$ and non-diagonal elements equal to zero, and denoting

$$\mu_i = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ u_{ip} \end{bmatrix}, \qquad i = 1, 2, \ldots, m,$$

we have

$$D(i,j) = \sum_{k=1}^{p} (\mu_{ik} - \mu_{jk})^2/\sigma_k^2. \quad (14)$$

To illustrate the use of divergence criterion for feature selection, consider the following example.

Example 2: Suppose we have three classes π_1, π_2, π_3 distributed normally with means $\mu_1 = (0, 0, 0)^T$, $\mu_2 = (0, 1, 0)^T$, $\mu_3 = (0, 0, 1)^T$, and common covariance matrix $\Sigma = I$. We want to select the best two characteristics by using the divergence principle. Using the formula in (14), Table 1 is obtained for the values of $D(i, j)$, i and $j = 1, 2, 3$, $i \neq j$.

Since the maximum of minimum divergences in different rows corresponds to pair (c_2, c_3), the divergence criterion leads to the selection of characteristics c_2, c_3. It is important to note that there is no discriminatory information provided by the characteristic c_1. Hence it is desirable that a feature selection procedure pick up c_2, c_3 if the best pair of features is required to be selected.

Taking the basic idea from Example 2, we establish a general result in the following Theorem 1, on the basis of which one can design a simulation model that can be used for evaluating a feature selection routine.

Theorem 1: Let the class π_i have the normal distribution with mean vector $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iq}, r_1, \ldots, r_{p-q})^T$ and covariance matrix $\Sigma = \text{diag}\ (\sigma_i^2,\ i = 1, 2, \ldots, p)$, $i = 1, 2, \ldots, m$. The class means are nonidentical for first q characteristics and identical for the remaining p–q characteristics. Then for the best selection with q characteristics a feature

Table 1. The values of $D(i, j)$

		Class pair		
		(π_1, π_2)	(π_1, π_3)	(π_2, π_3)
Characteristic pair	(c_1, c_2)	1	0	1
	(c_1, c_3)	0	1	1
	(c_2, c_3)	1	1	2

selection routine based upon the divergence criterion should lead to the selection of those characteristics for which the class means are non-identical.

Proof: Due to the given structure for class means, the divergence between classes π_i and π_j given by (14) reduces to

$$D(i,j) = \sum_{k=1}^{q} (\mu_{ik} - \mu_{jk})^2/\sigma_k^2, \tag{15}$$

i.e. any contribution in divergence is made by those characteristics which have non-identical class means. Suppose we consider an arbitrary set of q characteristics selected from the set of p characteristics such that t of these have non-identical class means and the remaining $q - t$ have identical class means, $t < q$. Then with the use of these t selected characteristics the divergence between π_i and π_j is equal to

$$\sum_{k_i=1}^{t} (\mu_{ik_i} - \mu_{jk_i})^2/\sigma_{k_i}^2$$

which is smaller than $D(i,j)$ in (15). Since this is true for any pair (π_i, π_j), the minimum divergence using first q characteristics over all possible pairs of classes exceeds that using any other set of q characteristics. Hence a feature selection routine based upon divergence principle should lead to selecting the q characteristics for which class means are non-identical. This establishes the theorem.

Theorem 1 can be generalized to the case where Σ is any arbitrary covariance matrix. However, if the covariance matrices are not the same, the above result may not necessarily hold.

Using the model of Theorem 1 for simulation purposes, one can investigate, and hence evaluate a feature selection algorithm by comparing the result of the characteristics it selects with the best possible selection of characteristics, the number of characteristics to be selected is preassigned.

5. CONCLUSION

Extensive development of the statistical methodology of classification has been stressed due to ever increasing demand for solving pattern recognition problems that often arise in many scientific disciplines, and especially in space science. Such need has already stretched the development from the mathematical treatment[6] of the problem to a technological level. In literature one now finds pure numerical classification techniques[13] as well as complicated software systems, involving clustering techniques, feature selection routines and discriminant algorithms [14, 15]. The purpose of this paper is to emphasize need for evaluating classification algorithms and to provide certain simulation models that can be used for verifying whether or not a classification system is performing as expected before being applied to a larger task involving a complicated situation.

Whether or not the system is useful depends on how it performs using unknown real data; hence, the concept of usefulness is highly application related. The approach discussed here will indicate whether or not the system performs as it was designed to perform and gives only indirect but important evidence as to its usefulness. If the system failed to perform as desired using simulated data, it would indeed be surprising if the system would be useful using real data: however, the converse is not true.

REFERENCES

1. P. L. Odell and R. S. Chhikara, Diagnostic Procedures for Simulation Evaluation of Some Discrimination Routines in Remote Sensing, Mimeographed Report, Texas Center for Research (1972).
2. W. G. Eppler, C. A. Helmke and R. N. Evans, Table look-up approach to pattern recognition, *Proc. 7th International Symposium on Remote Sensing of the Environment*, University of Michigan (1971).
3. P. L. Odell, B. S. Duran and W. A. Coberly, On the table look-up in discriminate analysis, *J. Statist. Comput. Simulation* **2**, 171–188 (1973).
4. M. D. Levine, Feature extraction: A survey, *Proc. IEEE* **57**(8), 1391–1408 (1969).
5. N. Glick, Sample-based classification procedures derived from density estimators, *J. Am. Statist. Assoc.* **67**, 116–122 (1972).
6. T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley, New York (1958).
7. G. Marsaglia, M. D. MacLaren and T. A. Bray, A fast procedure for generating normal random variables, *Commun. Assoc. Comput. Machinery* **7**, 4–10 (1964).
8. T. G. Newman and P. L. Odell, *The Generation of Random Variates*, Griffin's Statistical Monographs and Courses, No. 21 (1971).
9. S. Kullback, *Information Theory and Statistics*, Wiley, New York (1969).
10. T. Kailath, The divergence and Bhattacharya distance measures in signal selection, *IEEE Trans. Commun. Technol.* **15**(1), 52–60 (1967).
11. T. Marill and D. M. Green, On the effectiveness of receptors in recognition system, *IEEE Trans. Information Theory* (1963).
12. C. Y. Chang, Divergence and probability of misclassification, *NASA–EOD* 1970 (1971).
13. R. E. Haskell, A new program for the nonsupervised classification of multispectral data, *NASA–EOD* (1973).
14. D. A. Landgrebe and LARS Staff, LARSYAA, A Processing System for Airborne Earth Resource Data, LARS Information Note 091968, Purdue University, Lafayette, Indiana (1969).
15. M. Y. Su, M. C. Pooley and C. G. Hand, Statistical algorithms and computer programs for analysis of multi-spectral observations, *NASA CR*-103182 (1970).
16. P. L. Odell, Computational Problems Associated with Performing Discriminate Analysis in a Remote Sensing Application, Mimeographed Report, Texas Center for Research (1972).